Visualization Methods for the Exploration of High Dimensional Data

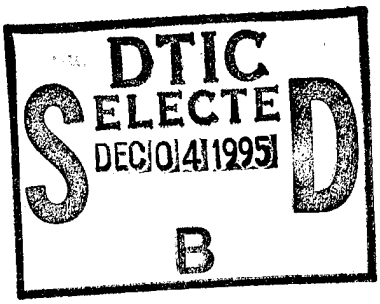
Final Report

Edward J. Wegman

February 16, 1995

U. S. Army Research Office

DAAL03-91-G-0039



George Mason University
Center for Computational Statistics

Introduction

This project focused on the development of data analytic and statistical methodology for data sets which may be characterized by one or more of the properties that they are large in size, high in dimension and nonhomogeneous. A major thrust is in the visualization of both data point clouds and mathematical structures in high dimensions. Several techniques were proposed including parallel coordinate density plots, 3-dimensional Andrews plots, grand (or guided) tours in 3 or higher dimensions. A combination of mathematical analysis and graphics display is our basic approach to these visualization problems. A closely related area is the area of structural inference for high dimensional structures. By this is meant the estimation of solid structures including k-dimensional flats in n-dimensional space as well as other nonlinear manifolds in high dimensional space. Proposed techniques involved 1. the detection and estimation of k-flats, thick k-flats and nonlinear manifolds of modest curvature by exploitation of the projective duality for parallel coordinates and 2. the estimation of more severely curved manifolds by use of ridges on k-dimensional density estimates. The parallel coordinate projective duality is that in parallel coordinates lines are represented by points and vice versa. Since k linearly independent lines are sufficient to uniquely specify a k-flat, it appeared to be possible to identify and arbitrarily oriented k-flat in n-space by appropriately exploiting parallel coordinates.

We proposed to focus on several aspects of computational statistics. The main focus was the development of methods for the visualization of multidimensional structure. The visualization of multidimensional structure is a key element in exploratory analysis of high dimensional data, but, of course, with much broader spinoff in terms of other scientific areas. We suggested four research topics related to the visualization: 1. Three-Dimensional Andrews and Related Plots, 2. The Grand Tour in Three Dimensions, 3. Finding Structure in k-Dimensions Using Grand Tour and Parallel Coordinates, and 4. Structural Inference using Ridge Estimation in Hyperspace.

Three-Dimensional Andrews and Related Plots

An Andrews plot is a multidimensional plotting device that is somewhat related to the parallel coordinate methodology. There are several conceptual viewpoints that can be described in connection with Andrews plots. First of all think of a data vector $(X_1,\ldots,X_n)$ as represented by pairs of the form $(1, X_1), \ldots, (n, X_n)$. One way of think of the parallel coordinate plot is as a linear interpolation between these points. The reason for using a linear interpolation is that the transformation from Cartesian space to parallel coordinate space is a projective transformation and, thus, leads to an elegant geometric interpretation of mathematical structure. In particular, we can map Cartesian geometric structures into parallel coordinate geometric structures. However, other general sets of interpolations may be suggested. The earliest one is essentially a Fourier interpolation. That is, plot a multidimensional vector as a trigonometric polynomial expansion with coefficients determined by the weights X_i. Specifically Andrews suggests plotting

$$f_x(\theta) = X_1/\sqrt{2} + X_2\ sin(\theta) + X_3 cos(\theta) + X_4 sin\ (2\theta) + X_5 cos(2\theta) + \cdots$$

Each unique point gets mapped into a unique trigonometric polynomial. These are then plotted in a way similar to parallel coordinate plots. Two properties of Andrews plots are interesting. First, because of the Fourier series interpretation, the classic Parseval's Theorem holds. Parseval's Theorem basically has to do with L_2-norms and asserts that mean square error in the Fourier domain and mean square error in the untransformed domain are the same. Thus while the untransformed domain is n-dimensional Euclidian space, the Fourier domain is 2-dimensional space so that by looking at an Andrews plot we can visually get an idea of the mean square error structure. The second property relates to the fact that we are talking about orthonormal trigonometric series. Because of this (thinking of the x-axis variable as an angle, say θ), for every θ we get a different linear weighting of the X_is. We can think of a slice at θ as a 1-dimensional projection of the multivariate vector onto an axis whose orientation is determined by θ. It has been argued that this in effect gives us a one-dimensional grand tour. As with any grand tour this offers us the possibility of looking for orientations that show up interesting or unusual properties.

There is nothing inherently sacred about either the piecewise linear (parallel coordinate plot) or the trigonometric (Andrews plot) interpolation. The former is useful because in preserves geometric properties, the

latter because of the mean square interpretation. The 1-dimensional grand tour would work with any orthonormal series so there may be some other interesting orthonormal series to think about. It may be that we can invent series which highlight different properties so that we can have a family of plots designed to explore different aspects of the structure. That is to say, if we are interested in highlighting clustering or outliers, we propose to invent an orthogonal series that would exaggerate those aspects of the data in the plot. Thus, we could generalize the parallel coordinate and Andrews plots.

Our work in this area was published in Wegman and Shen (1993) and also described in Wegman, Carr and Luo (1993), and Wegman and Carr (1993). One key result was to do an expansion in two dimensions instead of just one. What I have described before is an expansion $f(\theta; X)$ where $X=(X_1,\ldots,X_n)$ where f is either a piecewise linear interpolant or a trigonometric series. I used a bivariate expansion say $\vec{f}(\theta; X) = (f_1(\theta), f_2(\theta))$ as a 2-dimensional Fourier transform with irrational phase ratio (or, in fact, any orthonormal series). In this situation I was able to preserve the Parseval-type property and create the two-dimensional pseudo-grand tour. We can think of a 3-dimensional plot, plotting $\vec{f}(\theta)$ against θ. If the θ axis corresponds to the x axis and $\vec{f}$ to the y-z axis, we implemented this in our VR lab with rotation around the x-axis to help visualize the three-dimensional structure. Having a three-dimensional plot helps uncover more structure in the data than a simple two-dimensional plot would. Moreover, we are able to rotate the plot so that the y-z axis is the screen axis. Then slicing this graph along the x-axis would correspond to doing a two-dimensional grand tour. This provided a unified treatment of Andrews/parallel-coordinate-type plots with the grand tour idea.

The Grand Tour in Three and Four Dimensions

The grand tour is a very interesting idea first and primarily exposited by Asimov, and never really given its full due we believe because it is computationally intensive and technically fairly difficult. The intuitive idea underlying the grand tour is as its name suggests, if we want to investigate a data set we "look at it from all angles" much as if we were doing a grand tour of a geographic place we would try and look at it in all aspects. Thus, for example, if we are exploring a ten-dimensional data set, we would like to look at it from as many different perspectives as we could. The original mathematical implementation was as projections into two-dimensional planes (flats). The collection of all two-dimensional flats in an n-dimensional Euclidian space is called a Grassmanian manifold. The idea is to create a space filling path (i.e. one that visits all elements of the Grassmannian manifold) in some continuous fashion with the additional restriction that the proportion of time spent in each region is proportional to the size of that region. That is to say we do not linger in a small region of the whole space. If we then think of stepping along this path, we get a series of 2-dimensional planes onto which we can project the n-dimensional point cloud. If there is no structure in the point cloud, then every two-dimensional projection should look like an uncorrelated scatter diagram. If there is (two-dimensional) structure, then some projections will have interesting non-trivial patterns and these can be modeled. Two problems arise. First, if the dimension of the data space is high, then the number of two-flats needed to get a reasonably dense collection of two planes is very large. This means that in any real implementation there is a tradeoff between density of planes and reasonable viewing time. However, if the density of viewing planes is fairly low, some perspectives will be missed and consequently some interesting projections may be lost. Second, the methods for choosing the path through the Grassmannian manifold are either computationally very tedious or not mathematically elegant and visually unappealing. Moreover, even if these aspects could be dramatically improved, it is clear that looking a sequence of two-dimensional projections will allow us to detect unusual two-dimensional patterns, but it will not necessarily allow for us to detect unusual patterns in 3 dimensions.

The fact that we have 3-dimensional display devices suggests that we could and have tried creating a grand tour in three dimensions. The idea is in an n-dimensional space there would be a large number number of three-dimensional subspaces. Instead of stepping through a sequence of two-flats, we could step through a sequence of three-flats. There are, of course, as many two-dimensional flats (coordinate systems) as there are three-dimensional flats (coordinate systems) in the sense that both have the same cardinality and are uncountably infinite. Nonetheless, in a practical implementation, we do not have to step through as many 3-D coordinate systems as 2-D coordinate systems in order to densely approximate all possibly systems. In the two-dimensional grand tour we are interested in determining two-flats. These will be determined by a pair of unit length vectors, say **(a, b)**, which are orthogonal and which span a given plane. Of course, if each of the

components of (**a**, **b**) contain only 0s and 1s, these will correspond to planes of the original coordinate axes system. Thus the 2-flat of interest is span(**a**, **b**). We have achieved two important results: 1) We have generalized the grand tour to general k-dimensional representations, i.e. we have created a time-dependent series of orthonormal vectors in k-dimensions, $(\mathbf{a}_1(t), \mathbf{a}_2(t), \ldots, \mathbf{a}_k(t))$ (see description below) and, 2) We have found a computationally efficient algorithm for a 2-dimensional pseudo-grand tour (see description above). These results were reported in Wegman (1991b), Wegman and Shen (1993), Wegman, Carr and Luo (1993) and Wegman and Carr (1993).

Finding Structure in k-Dimensions using Grand Tour and Parallel Coordinates

The project here is conceptually closely related to our earlier discussions of the grand tour. As indicated earlier, the advantage of doing a 3-dimensional grand tour is two-fold. First, it allows for one to see unusual 3-dimensional configurations instead of simply unusual 2-dimensional configurations. Second, it allows a more complete search of the k-dimensional space because, for practical purposes, there are fewer 3-flats needed than 2-flats to attain the same density. Because parallel coordinates is a convenient tool for representing data in dimensions 4 and higher, a natural suggestion is to combine the parallel coordinate representation with the grand tour notion. Generalizing our earlier notion, suppose $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k)$ is a vector of k-unit vectors which form the mutually orthogonal unit vectors whose span is a k-flat. This is, so to speak, a Grassmannian manifold of k-flats instead of 2-flats. The idea is to find a continuous, dense path through this Grassmannian manifold and use the k-flats (k-dimensional coordinate system) so generated as a sequence of coordinate systems in which we can plot the data. Of course, we would not plot using Cartesian coordinates, but we can plot using parallel coordinates. Again we would be searching for unusual structure. One structure that would be of interest finding that the data lie on one or more k-flats or other k-manifolds. For example, verifying that the data were co-(k-)planar in some orientation of a k-flat would essentially suggest that a multiple linear regression with 1 dependent and $(k - 1)$ independent variables is an appropriate model. Other structures would suggest other models.

The trade-off is obvious. As k gets larger, the ability to look for unusual higher dimensional structure improves. Also the density of k-flats is much high than the density of 2- or 3-flats and so it appears plausible that we could look more closely at the n-dimensional space. The bad news is that the computation of the unit vectors $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k)$ is likely to become computationally more intensive. How bad this might be is not yet clear.

Let us consider a related observation. We know that lines in parallel coordinates represent points in Euclidian space and similarly, points in parallel coordinates represent lines in Euclidian space. Suppose we have a bunch of points in Euclidian space chosen randomly except that they all lie on a plane, say a d-flat. They are represented by a collection of line segments joining parallel coordinate axes. Let's let the ith point in Euclidian n-space be represented by $\mathcal{L}^i$ and let the line between axis j and $j + 1$ be $\mathcal{L}^i_j$, $j = 1, 2, \ldots, n - 1$. The intersection of $\mathcal{L}^i_j$ and $\mathcal{L}^k_j$ is a point in parallel coordinate space representing a line in Euclidian space denote it by $\mathcal{P}^{ik}_j$. Joining this to $\mathcal{P}^{ik}_{j+1}$ gives us a new line segment in parallel coordinate space, say $\mathcal{L}^{ik}_j$, which represents a point in Euclidian n-space. Since the lines represented by $\mathcal{P}^{ik}_j$ are coplanar, their intersections represented by $\mathcal{L}^{ik}_j$ are also coplanar. This implies that all of the segments $\mathcal{L}^{ik}_j$ should have a common intersections as j ranges from 1 to n. Indeed, if there is not one but several intersections for each j, this suggests that there are not one but several planes. Generalizing this process to higher dimensions this suggests another diagnostic tool for detecting when a point cloud lies on one or more k-flats. Coupled with the k-dimensional grand tour, this may be a very powerful geometric diagnostic tool for inferring data structure in higher dimensions.

A related problem is to diagnose nonlinear structure. If we have data on a nonlinear k-manifold, the given technique may not be entirely appropriate. This technique is fairly robust to variability to some scatter off of the plane (i.e., when dealing with a thick slab). If so, then a k-manifold which has small to moderate curvature may be regarded as a modestly thick slab and although the $\mathcal{L}^{ik}_j$ will not have exactly common intersections the intersections should cluster tightly. The idea is then to introduce nonlinear transformations of the data and look at the plot of the intersections as a graphical tool for diagnosing how well the transformation is linearizing the data fit. Of course, if the k-manifold is highly curved, there may not be any indication of

planarity. This work was reported in Wegman (1991b) and has been coded into a software package titled ***ExplorN*** which is co-authored by Carr, Luo, Wegman and Shen.

Structural Inference using Ridge Estimation in Hyperspace

This problem arises from an attempt to abstract the general idea of nonparametric regression. The idea of regression, of course, is that there is a response variable, say Y, and one or more predictor variables, say $X_1, \ldots, X_d$. In regression we attempt to find a function, say f, so that Y is approximated by $f(X_1, \ldots, X_d)$ in some sense, usually least squares. This gives the random variable Y some sort of preferred status over the variables $X_1, \ldots, X_d$. This may or may not be appropriate. We can however think of the variables $Y, X_1, \ldots, X_d$ as a vector which describe a point in a d+1 dimensional space. These points satisfy some functional relationship, that is there exists a function, say F, such that $F(Y, X_1, \ldots, X_d) = 0$. Another way of thinking about this is geometrically, i.e. $\mathfrak{M} = \{(Y, X_1, \ldots, X_d): F(Y, X_1, \ldots, X_d) = 0\}$ is some sort of hypersurface of dimension k embedded in a d dimensional space. To make this concrete by an example, let $d = 2$, $k = 1$ and $F(Y, X) = Y - \sin(X)$. Then the points (Y, X) in $\mathfrak{M}$ are exactly the points in two dimensions lying on the $Y = \sin(X)$ curve. $\mathfrak{M}$ is a one-dimensional set in a two-dimensional space. Because we are dealing with random variables we cannot expect the points to lie exactly on the hypersurface, $\mathfrak{M}$, (technically $\mathfrak{M}$ should be called an algebraic manifold), but to be scattered off of it. Thus we should really think about $F(Y, X_1, \ldots, X_d) = \epsilon$ so that taking expected values we find that $\mathfrak{M} = \{(Y, X_1, \ldots, X_d): E\, F(Y, X_1, \ldots, X_d) = 0\}$ is the manifold we would like to estimate. Notice in the regression case if we let $F(Y, X_1, \ldots, X_d) = Y - f(X_1, \ldots, X_d)$, then $F(Y, X_1, \ldots, X_d) = \epsilon$ corresponds to $Y = f(X_1, \ldots, X_d) + \epsilon$. In general, in this description we have left Y in to draw the analogy to usual regression, but Y is not intended to have a preferred status. Thus from now on we shall simply consider $F(X_1, \ldots, X_d)$ and define $\mathfrak{M}$ as $\{(X_1, \ldots, X_d): E\, F(X_1, \ldots, X_d) = 0\}$. Thus finding the functional relationship among the X's (i.e. in my language structural inference) is equivalent to estimating the manifold, $\mathfrak{M}$. Since $\mathfrak{M}$ is a geometric structure in hyperspace, we have the potential of visualizing it through some of our graphical techniques.

We suggest a connection with probability densities. Consider a plot of a two-dimensional normal density. In general this will be a surface in three dimensions. If we try to think of the best zero-dimensional summary of the density most people would probably suggest the mean. Since the mean and mode of the normal are co-located, this would also be the mode. Let me use language which suggests a solution for higher dimensions. The best zero-dimensional summary is location of mode which is the projection of the maximal zero-dimensional manifold on the surface of the density. If we try and think of the best one-dimensional summary, think of the fact that slices of the density parallel to the $X - Y$ have elliptical cross section with a major axis and a minor axis. Operationally we would probably want to choose our summary as the major axis of the density. Notice if the cross section were circular, correlation would be 1 and there would be no difference between major and minor axes. Basically it would not make sense to talk about a best one-dimensional summary. If, however, the correlation were plus or minus 1, the minor axis would have zero length and the major axis would coincide with the usual regression line. (Because of perfect correlation there would be no scatter off of this line.) If we think of the ridge on the density surface (ridge in the intuitive sense like on a mountain or hill), the major axis will lie beneath this ridge. In some sense, the ridge we have just described is the maximal one dimensional manifold on the surface of the two-dimensional density. The best 1-dimensional manifold estimate is the support of the ridge, i.e. the closure of the set of points for which the ridge is positive. The idea in general is to find the maximal k-dimensional manifold on the d-dimensional surface of the density which we will define as the k-ridge. The k-skeleton is k-dimensional manifold which is the support of the k-ridge in the d-dimensional space. The research problems was to construct a suitable definition of the k-ridge and to construct reasonable estimators. A potentially reasonable estimation procedure for the k-ridge is to estimate the probability density function and find the maximal k-ridge on it. Another element of the research was to implement a 3-dimensional surface projection of the k-skeleton for $k = 2$ or 3 either on the Silicon Graphics machine using our VR immersive technology. The 0-skeleton is the mode. These other estimators are multidimensional analogues of the mode. This work has been reported in Wegman, Carr and Luo (1993) and in numerous invited presentation. The completed research will form the substance of the dissertation of our Ph.D. student, Qiang Luo. Mr. Luo will be awarded his Ph.D. in May, 1995. The work has also been made available in software entitled, ***MasonRidge***, authored by Luo and Wegman.

Other Work

The four topic areas described above were the topics outlined in the research proposal upon which the award was made. However, there have been an extensive amount of additional work produced under this contract. This additional work generally falls into the categories of: 1) nonparametric density and function estimation (Le and Wegman, 1991; Miller and Wegman, 1991; Hearne and Wegman, 1991; Hearne and Wegman, 1992; Le and Wegman, 1993, Hearne, 1994; Marchette et al. 1994; Solka et al. 1994a; Hearne and Wegman, 1994 and Solka et al. 1994b), 2) parallel and high performance computing in statistics (Wegman, 1991a; Xu, Miller and Wegman, 1991; Sullivan and Wegman, 1994; Poston and Solka, 1994; Wegman and Jones, 1994; Takacs, Wegman and Wechsler, 1994; Fauntleroy and Wegman, 1994; Wegman, 1994; Sullivan, 1994; and Sullivan and Wegman, 1995), 3) stochastic modeling (Wegman and Habib, 1992; and Chow, 1994) and, finally, 4) historical (Wegman, 1992; Wegman, 1993).

Papers Published under ARO Sponsorship

Le, Hung Tri and Wegman, E. J. (1991), "Generalized function estimation of underwater transient signals," ***J. Acoust. Soc. America,*** 89, 274-279.

Miller, John J. and Wegman, E. J. (1991), "Construction of line densities for parallel coordinate plots," ***Computing and Graphics in Statistics***, (A. Buja and P. Tukey, eds.), 107-123, Springer-Verlag: New York.

Wegman, Edward J. (1991a), "A stochastic approach to load balancing in coarse grain parallel computers," in ***Computing and Graphics in Statistics***, (A. Buja and P. Tukey, eds.), 219-230, Springer-Verlag: New York.

Xu, Mingxian, Miller, John J. and Wegman, Edward J. (1991), "Parallelizing multiple linear regression for speed and redundancy: an empirical study," ***J. Statist. Comput. Simul.***, 39, 205-214.

Wegman, Edward J. (1991b), "The grand tour in k-dimensions," ***Computing Science and Statistics***, 22, 127-136.

Hearne, Leonard B. and Wegman, Edward J. (1991), "Adaptive probability density estimation in lower dimensions using random tessellations," ***Computing Science and Statistics***, 23, 241-245.

Wegman, Edward J. (1992), "Introduction to Box and Jenkins (1962) Some statistical aspects of adaptive optimization and control," ***Breakthroughs in Statistics, Volume II***, (S. Kotz and N. Johnson, eds.), 361-368, Springer-Verlag: New York, 1992

Wegman, Edward J. and Habib, Muhammad K. (1992), "Stochastic methods for neural systems," ***J. Statistical Planning and Inference,*** 33, 5-26.

Hearne, Leonard B. and Wegman, Edward J. (1992), "Maximum entropy density estimation using random tessellations," ***Computing Science and Statistics***, 24, 483-487

Le, Hung Tri and Wegman, Edward J. (1993), "A spectral representation for the class band-limited functions," ***Signal Processing***, 33(1), 35-44.

Wegman, Edward J., Carr, Daniel B. and Luo, Q. (1993), "Visualizing multivariate data," in ***Multivariate Analysis: Future Directions***, (Rao, C. R., ed.), Amsterdam: North Holland, 423-466.

Wegman, Edward J. and Carr, Daniel B. (1993), "Statistical graphics and visualization," in ***Handbook of Statistics 9: Computational Statistics***, (Rao, C. R., ed.), Amsterdam: North Holland, 857-958.

Wegman, Edward J. and Shen, Ji (1993), "Three-dimensional Andrews plots and the grand tour," ***Computing Science and Statistics***, 25, 284-288.

Wegman, Edward J. (1993), "History of the Interface since 1987: The corporate era," ***Computing Science and Statistics***, 25, 27-32.

Sullivan, Mark and Wegman, Edward J. (1995), "Correlation estimators based on simple nonlinear transformations," to appear ***IEEE Trans. Signal Processing***.

Technical Reports Prepared under ARO Sponsorship not listed above
(All are published by the Center for Computational Statistics at George Mason University)

TR 92. Leonard B. Hearne, Probability Density Estimation on a High Dimensional Space Using Random Tessellations (Ph.D. Dissertation), February, 1994.

TR 93. Winston C. Chow, Modeling and Estimation with Fractional Brownian Motion and Fractional Gaussian Noise (Ph.D. Dissertation), February, 1994.

TR 95. Mark C. Sullivan and Edward J. Wegman, Normalized Correlation Estimators Based on Simple Nonlinear Transformations, March, 1994.

TR 97. Wendy L. Poston and Jeffrey L. Solka, A Parallel Method to Maximize the Fisher Information Matrix, June, 1994 (appeared in Proceedings of Intel Supercomputer User's Conference, June, 1994).

TR 98. Edward J. Wegman and Charles A. Jones, Simulating a Multi-target Acoustic Array on the Intel Paragon, June, 1994 (appeared in Proceedings of Intel Supercomputer User's Conference, June, 1994).

TR 99. Barnabas Takacs, Edward J. Wegman and Harry Wechsler, Parallel Simulation of an Active Vision Model, June, 1994 (appeared in Proceedings of Intel Supercomputer User's Conference, June, 1994).

TR 102. Julia Corbin Fauntleroy and Edward J. Wegman, Parallelizing Locally-Weighted Regression, October, 1994.

TR 104. David J. Marchette, Carey E. Priebe, George W. Rogers and Jeffrey L. Solka, Filtered Kernel Density Estimation, October, 1994 (appeared in ***Computing Science and Statistics***, 26).

TR 105. Jeffrey L. Solka, Edward J. Wegman, Carey E. Priebe, Wendy L. Poston and George W. Rogers. A Method to Determine the Structure of an Unknown Mixture Using the Akaike Information Criterion and the Bootstrap, October, 1994a (tentatively accepted for ***Statistics and Computing***).

TR 106. Wendy L. Poston, Edward J. Wegman, Carey E. Priebe and Jeffrey L. Solka, A Contribution to the Theory of Robust Estimation of Multivariate Location and Shape: EID, October, 1994.

TR 109. Leonard B. Hearne and Edward J. Wegman, Fast Multidimensional Density Estimation based on Random-width Bins, October, 1994 (appeared in ***Computing Science and Statistics***, 26).

TR 110. Edward J. Wegman, Huge Data Sets and the Frontiers of Computational Feasibility, November, 1994.

TR 112. Mark C. Sullivan, Computationally Efficient Statistical Signal Processing Using Nonlinear Operators (Ph.D. Dissertation), December, 1994.

TR 114. Jeffrey L. Solka, Wendy L. Poston and Edward J. Wegman, A New Visualization Technique to Study the Time Evolution of Finite and Adaptive Mixture Estimators, December, 1994b (tentatively accepted in the ***Journal of Computational and Graphical Statistics***).

Personnel Directly Supported by the Contract

Edward J. Wegman, Principal Investigator
Daniel B. Carr, Associate Investigator
Don R. Faxon, in candidacy for Ph.D.
Farid Haq, M.S. awarded May, 1994
Hiroko Kawasaki, M.S. awarded May, 1993
Fan Zhang, in candidacy for Ph.D.
Sundari Chander, Ph.D. student
Ann Marie Clark, M.S. student
Charles Jones, M.S. awarded May, 1994
Mary Mortlock, M.S. awarded December, 1994
Rashmi Tandon, M.S. awarded December, 1994
James Turtora, Ph.D. student
Leonard Hearne, Ph.D. awarded May, 1994

Other Degrees Awarded to Students of Wegman
(Wegman's work supported by ARO)

Carey E. Priebe, Ph.D. awarded May, 1993
Osama A. Morad, Ph.D. awarded May, 1993
Celesta G. Bell, Ph.D. awarded May, 1993
Winston Chow, Ph.D. awarded May, 1994
Mark Sullivan, Ph.D. awarded December, 1994
Jeffrey Solka, Ph.D. to be awarded May, 1995
Wendy Poston, Ph.D. to be awarded May, 1995
Qiang Luo, Ph.D. to be awarded May, 1995
Cajetan Akujuobi, Ph.D. to be awarded May, 1995

MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	February 16, 1995	

4. TITLE AND SUBTITLE

Visualization Methods for the Exploration of High Dimensional Data

5. FUNDING NUMBERS

6. AUTHOR(S)

Edward J. Wegman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

George Mason University
Center for Computational Statistics
MS 4A7
4400 University Drive, Fairfax, VA 22030-4444

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This project focused on the development of data nalysis and statistical methodology which may be characterized by one or more of the properties that they are large in size, high in dimension, and nonhomogeneous. The major thrust is in visualization of both point clouds and mathematical structures in high dimensions. The specific advances made under this project are 1) Thre dimensioal generalizations of the Andrews plot, 2) the grand tour in k-dimensions, 3) fast algorithms for the pseudo-Grand Tour. and 4) structural inference using ridge estimation.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Andrews plot. grand tour, fast algorithms, ridge estimation	9
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- **DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- **DOE** - See authorities.
- **NASA** - See Handbook NHB 2200.2.
- **NTIS** - Leave blank.

Block 12b. Distribution Code.

- **DOD** - Leave blank.
- **DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- **NASA** - Leave blank.
- **NTIS** - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.